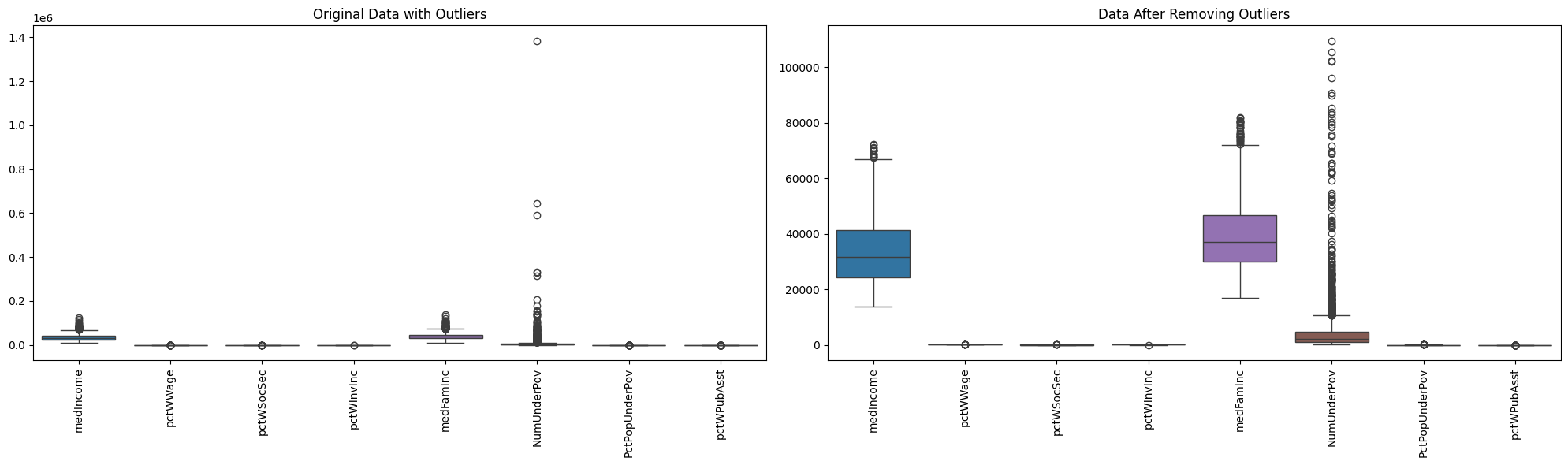
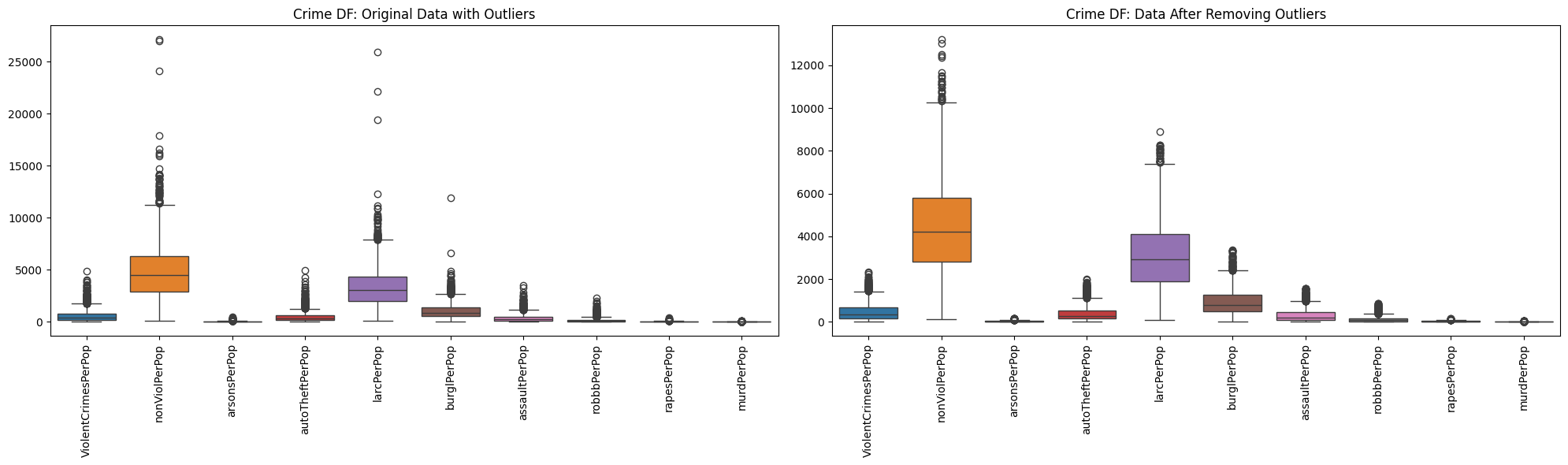
**Sprint 2: Data Preparation and Exploration**

Prior to creating models and exploring our data set it is important to ensure that the data set itself is properly prepared and cleaned. The first step taken to ensure this was to look into the data set and determine if there are any null values present in the data, null values can cause issues down the line when we begin to model the data so therefore they must be handled appropriately. We checked if any null values are present using the .isna() function on python, this function allows us to quickly check to see if any columns inside the data set contain null values. Upon executing this function it appeared that there were no actual null values present inside the data set, although upon further inspection it became apparent that although no null values were spotted while using the .isna() function this was because the way this data set dealt with null values was by assigning those values as a string ‘?’ and therefore they were not initially spotted. Our next step was to then do a similar process as previously but now instead of looking for null values we are looking for ‘?’ values. When checking for ‘?’ we discovered that 2104 out of the 2215 rows contained ‘?’, although at first glance this is 95% of our data we decided that we will first break down the data sets into subsets of variable we plan on exploring and then rechecked these subset data sets to determine how many ‘?’ values were present.

The two new data frames we created were one that contained all the crime variables and then another that contained all the income variables as these two subsets of the data set contained the variables we plan to use further in this project. Upon checking these two new data frames for ‘?’ values it revealed that our crimes subset contained 313 rows with null values and our income subset actually contained no null values at all so did not need to be altered. We decided to deal with the null values in the crimes subset by simply dropping these rows, our reasoning behind dropping these rows was because it was only a small portion of the data (roughly 14%) and also because this data pertains to specific state and counties crime rates and therefore we did not feel it appropriate to fill these null values with some sort of pseudo data.

Next step in our data preparation process was to then explore our two data frames to make sure that no outliers are present. The reason we need to ensure no outliers are present is because the presence of outliers could negatively affect our modeling and potentially skew hypothesis testing due to having inaccurate data. To get a better understanding of where potential outliers might be present we decided to visualize both of these new data frames using boxplots that would better reveal any outliers. In our boxplot for crime it was evident that the categories nonViolPerPop, larcPerPop and burglPerPop had the most obvious outliers, while in our income data set most categories seemed relatively clear of outliers except for the NumUnderPov category which had some very extreme outliers present. Our approach to cleaning up the outliers was to use z-score in order to determine whether outliers are present or not. Z-score is a measure of how far a data point is from the mean of the data set with respect to standard deviation. The formula for z-score is as follows: z = (X−μ)​ / σ, with X representing the data point, μ representing the mean of the data set, and σ representing the standard deviation of the data set. The goal of using the z-score is to determine whether the data point being checked is greater than 3 or less than -3, if so this indicates that the data point is clearly different from the other data points and therefore would be dropped due to being an outlier. Upon conducting z-scores on both our crimes and income dataset we determined there to be 164 outliers to be removed in the crimes data frame and 111 outliers to be removed in the income data frame. To better understand the following visualizations where created.

****

Finally the last bit of preprocessing we had done on our data was remapping the PctNotHSGrad column. The way we did this was by instead of having the values of this field be float values we decided to give them categorical values instead. This was done by splitting them into groups based on their graduation rates such as 0%-10%, 11%-20% and so on and so forth. Our reasoning behind this is that by splitting this category into categorical groups we can better utilize different visualizations to get a better understanding of our data. The float values of this column will still be used in the modeling process but this categorical remapping will be used as well.

This leaves us with almost entirely numeric fields that we will be working with. For the vast majority of our variables will be float values, although a few integer values exist inside our income variables. The only non-numerical field we have is the PctNotHSGrad that was broken down into categorical values rather than numerical.

Now that our data has been properly prepared and cleaned, we then determined our target and predictor variables. For our target variable we have decided to use violent crime rates as this is what we are trying to predict a decrease in and therefore it will be our target variable. In order to predict this target variable we will use the following predictor variables: percent of student without a high school degree, medium income, percentage with wage, percentage of people with social security, percentage of people with investment income, median household income, number of people under the poverty line, percentage of population living under the poverty line, and percentage of people acquiring public assistance. The reason we chose these predictor variables is because our goal is to emphasize the importance these variables have on violent crime, more specifically we believe that these predictor variables all work together to emphasize our notion that increasing the percentage of high school graduates will in term decrease the levels of violent crimes in our society. The income variables were chosen because we believe they are also directly related with high school graduation rates and could be used tangentially to further emphasize our hypothesis.